多级计分测验中基于残差统计量的被试拟合研究*

童昊[1] 喻晓锋[1] 秦春影[2] 彭亚风[1] 钟小缘[1]

([1]江西师范大学心理学院，南昌，330022)
([2]南昌师范学院数学与信息科学学院，南昌，330032)

摘要：本文提出一种多级计分项目下的个人拟合统计量R，考察它在检测 6 种常见的异常作答模式(作弊、猜测、随机、粗心、创新作答、混合异常)下的表现，并与标准化对数似然统计量l_{zp}进行比较。结果表明：(1) 在异常作答覆盖率较低并且异常作答类型为作弊和猜测时，R的检测率显著高于l_{zp}；(2) 随着测验长度和被试异常程度的增加，两种统计量的检测率都会上升；(3) 在一些条件下，R与l_{zp}检测效果接近。实证数据分析进一步展示了R统计量的使用方法和过程，结果也表明R统计量具有较好的应用前景。

关键词：多级计分项目，项目反应理论，个人拟合统计量，异常行为检测，等级反应模型

1 引言

教育和心理测量的主要目的是获得被试的某种潜在特质，来评估和指导个体的未来发展活动。该特质可以是学科领域内知识和技能的掌握情况，也可以是个体的态度、情绪等。为此，研究者广泛使用测验或问卷作为测量潜在特质的手段。然而在实际施测过程中，几乎无法避免地会有其他额外因素影响被试的作答反应，进而威胁测验结果的有效性，如作弊、低测验动机等。不同于被试作答过程的随机误差，这些因素往往能够导致测量数据产生系统误差，这类被试作答也被称为异常作答。

经典测量理论(classical test theory, CTT)和项目反应理论(item response theory, IRT)是常见用以估计被试特质水平的理论手段，其准确性依赖于模型和数据的拟合程度(Hotaka & Maeda, 2017)，若直接使用存在异常的作答数据进行分析和计算，得到的结果将产生较大的偏差，由此推导出的结论和基于结论做出的任何操作(人才选拔和岗位安置等)也将不再具有参考价值，对测验的很多方面产生严重的负面影响，比如参数估计(Oshima, 1994; Schnipke, 1996; Shao et al., 2016)、等值 (Wollack et al., 2003)、信度和效度(Gulliksen, 1950, p.

收稿日期: 2021-08-05
* 全国教育科学规划项目(BGA210060); 江西省社会科学基金项目(21JY06); 江西省高校人文社会科学项目(XL20202); 南昌市教育大数据智能技术重点实验室(2020-NCZDSY-012); 江西省教育厅科技项目(GJJ191691, GJJ191128)资助。
通信作者: 喻晓锋, E-mail: xyu6@jxnu.edu.cn

236; Lu & Sireci, 2007)等，进而最终导致测验的公平性和准确性受到损害(Buchanan & Smith, 1999; Glas & Dagohoy, 2007; Huang et al., 2015)。许多研究对于异常作答在考生中的“流行程度”进行了报告，比如 Curran 等人(2010)， Meade 和 Craig (2012)，Rupp (2013)，Meade (2016)，Shao 等(2016) ，Yu 和 Cheng (2019)等，这些研究中的异常作答人数占比从最低的 3.5%到最高的 50%，以中等程度 20%居多。因此，如何将存在异常作答模式的个体筛选出来，一直是教育和心理测量领域所广泛关注的问题 (Schnipke & Scrams, 1997)。

为了达到上述目的，研究者们通过构建个人拟合统计量(person fit statistic, PFS)从数据中挖掘被试信息，从而判断是否为异常。Meijer 和 Sijtsma(2001)将 PFS 划分为两类：(1) 非参数化 PFS：使用被试的观察作答数据计算非参数统计量或将个体与团体进行比较；(2) 参数化的 PFS：设定基本的模型假设，利用数据进行参数估计，构建拟合统计量，与零假设(受检验个体属于正常群体)下的统计量分布相比较，判断数据和模型拟合与否。

关于个人拟合的研究最早可以追溯到 20 世纪 40 年代，Guttman (1944, 1950)综合被试的观察作答、估计能力、项目截断点(cutting point)之间的关系，来判断其作答是否正常，这为之后的个人拟合研究奠定了基础。后续有研究者提出非参数化的 PFS 如点二列相关和二列相关(Donlon & Fischer, 1968)等，虽然这类非参数化检验方法在使用上较为便捷，仅需对原始数据进行少量的计算，但通常难以深度挖掘数据中的更多信息，对计算结果的解释不够明确。相比之下，参数化的 PFS 依靠严格的数学模型(如 Rasch 模型等)，在一定的数据规模下，量化被试的异常程度，依照设定的严格判别标准，得出清晰的结论。进一步，参数化 PFS 可以粗略分为基于残差和基于似然两种类型。前者的主体为观察作答和期望作答之间的残差，其建模思路是对残差进行适当加权处理，如 Wright 和 Stone(1979)以及 Wright 和 Master(1982)提出的U和W，分别使用了单个项目的平均条件方差和测验总方差的倒数作为加权内容。后者则是围绕作答的似然来构建统计量，例如 Levine 和 Rubin(1979)提出的对数似然指标l_0，以及 Drasgow 等人(1985)克服了l_0分布缺陷而提出的标准化对数似然指标l_z。

当前，个人拟合检验的研究主要集中在二级(0-1)计分的背景下，然而在实际教育和心理评估测验中存在大量多级计分的数据，例如，混合测验中的主观题，或心理测验中常使用的李克特型(Likert-type)量表问卷。与二级计分的项目相比，多级计分项目能够提供更多的信息，只需要更少的题目就能达到和较多二级计分项目同样的测量精度(van der Ark, 2001)。多级计分下的参数化 PFS 主要有 Wright 和 Masters(1982)提出的标准化加权均方残差统计量υ，以及 Drasgow 等(1985)提出的多级标准化对数似然指标l_{zp}。但以上两种个人拟合

统计指标存在一些缺陷：(1)Rogers 和 Hattie(1987)指出υ对异常作答模式的分类并不敏感，本研究在预实验阶段也证明了其较差的检测能力，因此υ的实用价值十分有限。(2)对于l_{zp}而言，它拥有较好的标准化分布形态，和较为优秀的综合检验效果，但是在实际应用中，研究者往往更关心低能力者的异常高能力表现，这是因为高风险测验的结果往往能够带来利害影响，被试的异常高能力表现通常意味着发生了诸如试题泄露、考生作弊等较为严重的测验安全事故；与之对应的，高能力者的异常低能力表现一般是由于被试个人原因如疲劳、加速作答等个体因素造成的。因此，为了更好地维护测验安全，我们亟需一种针对异常高能力敏感的多级计分 PFS。

2 多级计分模型及拟合统计量

2.1 多级计分 IRT 模型

研究者提出了多种用以处理多级计分数据的 IRT 模型，如等级反应模型 GRM(graded response model; Samejima, 1969)和分部评分模型 PCM(partial credit model; Masters, 1982)等。本文使用的是 GRM，其研究结果同样可以推广到其他模型如 PCM 等多级计分模型中，研究者可以根据应用情景进行选择。

在 GRM 中，单个项目包含一个区分度参数和多个等级难度参数。定义X_j表示被试在项目j上的作答，K表示项目j的满分值，则$X_j \in \{0,1,\ldots,K\}$，用θ表示被试的潜在特质水平，a_j和$\boldsymbol{b_j}$分别表示项目的区分度和难度，且$\boldsymbol{b_j}=(b_{j1},b_{j2},\ldots,b_{jK})$。对于每一个难度$b_{jk}$，被试都存在一定概率$P_{jk}^*(\theta)$获得$k$及以上的评分，此时可以用二参数逻辑斯蒂克模型进行描述：

$$P_{jk}^*(\theta)=\frac{1}{1+e^{-1.702a_j(\theta-b_{jk})}} \tag{1}$$

由于单个P_{jk}^*只能表示获得该评分等级及以上的概率，为了细化每一个评分等级的概率，需要将相邻难度对应的概率值P_{jk}^*与P_{jk-1}^*相减，可以得到被试在项目j上恰好获得k评分等级的概率：

$$P_{jk}=P_{jk-1}^*-P_{jk}^* \tag{2}$$

特别地，对于第一个评分等级，被试获得该评分的概率$P_{j1}=1-P_{j1}^*$。

假设某项目的满分为 4 分，即有 5 个评分等级(0,1,2,3,4)，此时，被试获得每个评分等级的概率如下所示：

$$\begin{cases} P_{j1}^*(X_j=0|\theta)=1-\dfrac{1}{1+e^{-1.702a_j(\theta-b_{j1})}} \\ P_{j2}^*(X_j=1|\theta)=\dfrac{1}{1+e^{-1.702a_j(\theta-b_{j1})}}-\dfrac{1}{1+e^{-1.702a_j(\theta-b_{j2})}} \\ P_{j3}^*(X_j=2|\theta)=\dfrac{1}{1+e^{-1.702a_j(\theta-b_{j2})}}-\dfrac{1}{1+e^{-1.702a_j(\theta-b_{j3})}} \\ P_{j4}^*(X_j=3|\theta)=\dfrac{1}{1+e^{-1.702a_j(\theta-b_{j3})}}-\dfrac{1}{1+e^{-1.702a_j(\theta-b_{j4})}} \\ P_{j5}^*(X_j=4|\theta)=\dfrac{1}{1+e^{-1.702a_j(\theta-b_{j4})}} \end{cases} \tag{3}$$

2.2 υ统计量

Wright 和 Master(1982)提出了一种标准化加权均方残差统计量υ，其表达式如下：

$$\upsilon=\frac{\sum_{j=1}^{M}\left[X_j-\sum_{k=0}^{K}k*P_{jk}(\theta)\right]^2}{\sum_{j=1}^{M}\left[\sum_{k=0}^{K}\left(k-E\left(X_j|\theta\right)\right)^2P_{jk}(\theta)\right]} \tag{4}$$

$$E\left(X_j|\theta\right)=\sum_{k=0}^{K}k*P_{jk}(\theta) \tag{5}$$

其中，M为项目个数，K为满分值，X_j表示被试在项目j上的得分，$P_{jk}(\theta)$表示能力为θ的被试在项目j上恰好获得k评分的概率，$E\left(X_j|\theta\right)$表示被试的期望得分。

在本文所关注的异常作答条件下，我们的预实验表明υ缺乏足够的检验性能(见附录 A)，因此后续模拟研究不考虑使用υ作为比较对象。

2.3 l_z统计量

对数似然统计量l_0最早是在二级计分下提出(Levine & Rubin, 1979)，多级计分下，Drasgow 等(1985)在l_0的基础上进行了标准化处理，得到标准化对数似然统计量l_z，并且推导了适用于多级计分的l_{zp}。对于某被试在测验上的作答反应模式$\boldsymbol{X}=\{X_1,X_2,\dots,X_M\}$，对数似然函数$l_p$的表达式如下：

$$l_p=\ln[L(\theta)]=\sum_{j=1}^{M}\sum_{k=0}^{K}d\left(X_j=k\right)P_{jk}(\theta) \tag{7}$$

$L(\theta)$为似然函数，$d(\cdot)$为指示函数，表示当满足括号内条件时取 1，否则为 0。X_j表示被试在项目 j 上的观测得分，$P_{jk}(\theta)$表示被试在项目j上恰好获得k评分等级的概率。对l_p进行标准化可以得到：

$$l_{zp}=\frac{l_p-E\left(l_p\right)}{\sqrt{Var\left(l_p\right)}} \tag{8}$$

其中，$E\left(l_p\right)$和$Var\left(l_p\right)$分别表示l_p的期望和方差，表达式如下：

$$E(l_p) = \sum_{j=1}^{M}\sum_{k=0}^{K} P_{jk}(\theta) \ln P_{ijk}(\theta) \tag{9}$$

$$Var(l_p) = \sum_{j=1}^{M}\left[\sum_{k=0}^{K}\sum_{h=0}^{K} P_{jk}(\theta) P_{jh}(\theta) \ln P_{jk}(\theta) \ln \frac{P_{jk}(\theta)}{P_{jh}(\theta)}\right] \tag{10}$$

l_{zp}是对l_p的标准化结果，具有良好的分布形态(渐进正态分布)，相较于大部分 PFS，各类实验条件下都能够拥有较好的检验性能(Karabatsos, 2003)，Nering (1995)将其称为“最具前景的 PFS”。

3 基于加权残差的多级计分 PFS 开发

使用模拟实验对已有的 PFS 进行性能比较时，我们发现，基于残差的 PFS 在异常高能力(作弊、幸运猜测)检测中具备一定优势，这与 Karabatsos(2003)的研究结果一致。为了解释该种现象，我们分析了不同 PFS 在多级计分背景下的构建思路，结果表明，基于残差的 PFS 具有对更高评分等级的敏感性。基于此，本研究在多级计分 IRT 框架下，对基于残差的参数化统计量进行拓展，开发了一种对异常高能力敏感的个人拟合统计量R。

构建基于残差拟合统计量的中心思想是量化观察反应模式和理想反应模式之间的差异。理想反应模式是指给定了各模型参数(如能力、难度、区分度等)后，严格依据反应函数的概率分布产生的反应模式，被试会更容易在相对自身高难度的项目上获得低分，反之在相对自身低难度的项目上获得高分。不拟合的作答模式必然会与理想反应模式之间存在较大的偏差(Meijer & Sijtsma, 2001)，这种偏差可以作为判断异常的依据。为了更准确地计算残差，使用期望得分代表理想反应。令X_j表示被试在项目j上的观察作答，$E(X_j|\theta)$为期望得分，此时被试的单个项目得分残差$r_j = X_j - E(X_j|\theta)$，根据公式(5)可以进一步分解为：

$$r_j = X_j - E(X_j|\theta) = \sum_{k=0}^{K} k * d(X_j = k) - \sum_{k=0}^{K} k * P_{jk}(\theta) =$$

$$\sum_{k=0}^{K} k * [d(X_j = k) - P_{jk}(\theta)] \tag{11}$$

这里K表示项目j的满分值，$d(\cdot)$为指示函数，满足条件时取 1，否则为 0。$P_{jk}(\theta)$表示被试在项目j上恰好获得k评分等级的概率。对单个项目的残差而言，可以拆分为k计分等级下，指示函数与概率差值的加权求和。观察公式(10)发现，每个计分等级差值的权重是得分k，所以随着k的增大，该项的权也就越大。正常情况下，被试的作答应该接近概率分布中最高的几个计分等级，对应的残差求和应该偏小；而对于存在异常的被试作答，残差将会显

著偏高，特别是异常高能力作答，高的计分等级赋予了该部分残差更大的权重。因此，基于残差的统计量在构建基础上具有对异常高能力表现的敏感性。

但是仅考虑残差本身，并不足以体现每个项目之间可能存在的差异，例如对同一个被试而言，在不同参数的项目上应具有不同的得分概率分布。因此，即使两个项目上得分的残差相同，也不应该将二者划上等号。参考 Snijders(2001)的研究，一个常见的做法是对残差添加权重函数$w_j(\theta)$，它能够综合项目和被试特点，让残差项在统计量中具有更合理的贡献。

权重函数的基本思路是放大“可疑”的部分，缩小相对正常的部分，以此来实现对异常的高敏感性。例如，被试在项目上获得了与期望得分接近的分数，此时的加权函数应当略小，才能将正常作答数据对统计量的贡献降低；反之，如果被试在项目上获得了显著偏离期望得分的分数，这部分数据的异常影响应当在统计量中得到放大。因此，参考 Master 和 Wright(1997)使用的标准化残差和 Yu 和 Cheng(2019)使用的加权残差，将权重函数定义如下：

$$w_j(\theta) = \left(\sum_{k=0}^{K} d\left(X_j = k\right) * P_{jk}(\theta) \right)^{-1} \quad (12)$$

这样的设置下，权重函数的大小取决于实际得分的理论概率值。此概率越大，说明被试的作答是相对正常的，因为其遵循了得分概率分布；概率越小，说明被试的作答越异常。所以，$w_j(\theta)$在正常情况下偏小，在异常情况下偏大。

另外，加权残差统计量是通过累加全体项目来体现被试的异常程度。由于被试在测验中可能存在混合类型的异常作答行为，如在部分项目上作弊，同时在测验尾部加速作答。导致高能力表现的异常项目残差和低能力表现的异常项目残差相互抵消，在一定程度上影响检测力。为了避免这种情况，最大限度发挥统计量的检测效果，需要对异常进行积累处理。

在实际应用中，被试的真实能力值θ往往无从得知，需要用估计值$\hat{\theta}$替代。对于存在异常的被试，$\hat{\theta}$和$E\left(X_j|\hat{\theta}\right)$偏离的方向相同，这会导致计算时的残差必然小于理论值，例如，某低能力被试($\theta = -2$)在测验中作弊，导致高估其能力($\hat{\theta} = 1$)，在该被试涉及作弊的某个项目上，观测得分$X_j = K$，由于$E\left(X_j|\hat{\theta}\right) > E(X_j|\theta)$，则$\left[X_j - E\left(X_j|\hat{\theta}\right)\right] < \left[X_j - E\left(X_j|\theta\right)\right]$。这对统计量的检验效果是不利的，因此，本研究采用取绝对值来实现残差的积累。定义R的表达式如下：

$$R=\sum_{j=1}^{M}|r_j|*w_j(\theta)=\sum_{j=1}^{M}\frac{\left|\sum_{k=0}^{K}k*\left[d\left(X_j=k\right)-P_{jk}(\theta)\right]\right|}{\sum_{k=0}^{K}d\left(X_j=k\right)*P_{jk}(\theta)} \tag{13}$$

由于计算出的期望得分往往是非整数，所以即使是完全正常的被试，其在测验上的残差累积和也几乎不可能为零。在理想情况下，R将会控制在一个较小的范围内，相对应的，异常反应模式下的R将会显著偏离零分布(零假设条件下的统计学分布)。这也是依靠 PFS 检验异常被试的依据——在获得了R的零分布后，通过设置不同一类错误率的截断点，来判断受检验被试的R是否处于接受域内。由于R是加权残差取绝对值后的累加和，其大小体现了作答模式的拟合程度，所以在进行假设检验时，使用的是分布右侧的单侧概率。需要注意的是，当考生的观测得分中存在异常数据时，基于此观测得分进行项目参数和被试参数的联合估计，会出现“掩蔽效应” (masking effect; Fung, 1993; Yuan & Zhong, 2008)，即异常数据会在一定程度上影响参数估计，进而降低其被检测出的可能性。这不利于比较R和l_{zp}的表现。因此，本研究中的模拟实验考虑在已知项目参数时，比较它们的表现。

4 模拟实验

模拟研究的主要目的是为了检验R对异常被试群体的检测能力，为此我们需要一个相对良好的比较对象，来直观地展现R的优势和特点。鉴于l_{zp}拥有优良的综合性能，受到研究者的广泛关注 (de la Torre & Deng, 2008; Sinharay, 2016)。故本研究选取l_{zp}作为比较对象，并以此开展了三项研究：研究一在零假设条件下，模拟正常作答群体，获得R和l_{zp}于不同测验长度下的统计学分布，同时为后续的研究二提供了基础；研究二通过模拟被试可能存在的不同异常测验情境，使用R和l_{zp}对数据进行检测，并基于虚警率和检测率来评价；研究三中将R应用于实证数据分析，观察其表现。

使用 PFS 进行拟合检验的基础是获取零分布，对于l_{zp}等经过标准化的 PFS，仍然需要谨慎对待。Sinharay(2016)发现，$l_{zp}(\hat{\theta})$并非服从渐进标准正态分布； van Krimpen-Stoop 和 Meijer(2002)的研究表明，当测验长度较小时(如 20,30)，l_{zp}的分布呈负偏态分布。考虑到实际情境中使用的是被试能力估计值$\hat{\theta}$,，因此可以认为l_{zp}并不服从标准正态分布。在l_{zp}截断点的选取上，不能直接照搬标准正态分布下的特定值。为了探寻和比较R和l_{zp}在多级计分测验中的异常探测能力，本文设计以下模拟研究：研究一，得到R和l_{zp}的分布和临界值。研究二，计算和比较R和l_{zp}的虚警率与检测率。

4.1 研究一：R和l_{zp}的分布及临界值

本研究使用的多级计分模型是 GRM，参考许多已有多级计分的研究(比如陈青等, 2010;

程小扬等人, 2012; Dodd et al., 1995; Emons, 2008; 李佳, 丁树良, 2018; Sinharay, 2016)，使用的记分等级均为五等。构建四种长度条件下的测验，包含 20、40、60、80 个项目，分别代表短、中、较长、长测验。单个项目内的难度参数从标准正态分布中选取，并按照升序排列；区分度参数从均值为 0，标准差为 1 的对数正态分布中选取 (Xiong et al., 2020; 熊建华等, 2018)。每个长度条件下，模拟 10000 名能力值θ服从标准正态分布的被试参加测验。根据 GRM 得到模拟被试的得分分布，生成观测得分。

计算l_{zp}和R的分布时，本研究基于已知项目参数，以及被试能力的估计值$\hat{\theta}$。这是因为在实际测验中，项目一般经过了精心的编制和多次的施测，可以认为，项目参数得到了校准，是已知的，这与 Shao 等(2016)和 Sinharay(2016)的做法相同。另外，由于项目参数已知，因此一类错误率和检验力不会随着待测数据样本中的被试人数的改变而受影响(Sinharay, 2016)。为了将研究结果更好地推广到真实的测验情境中，本研究使用能力估计值$\hat{\theta}$来进行后续的研究，估计方法为期望后验估计(expected a posteriori, EAP)，其表达式如下：

$$\hat{\theta} = \frac{\int \theta L(\theta) f(\theta) d\theta}{\int L(\theta) f(\theta) d\theta} \quad (14)$$

$L(\theta)$为似然函数，$f(\theta)$为θ的概率分布密度函数。相较于极大似然估计(maximum likelihood estimate, MLE)，EAP 利用了被试群体的先验信息，在一定程度上能够提高估计精度。当先验信息不明确时，可以适当放宽先验分布，采用部分信息先验或改用 MLE。

通过l_{zp}和R的定义可知，二者均在单侧反映拟合程度，区别在于l_{zp}越小说明越不拟合，R越大说明越不拟合。因此在截断点的选取上，设置一类错误率分别为 1%、2.5%、 5%，选取了l_{zp}分布的第 1、2.5、5 百分位数和R分布的第 99、97.5、95 百分位数。

图 1 和图 2 分别给出了R和l_{zp}的经验分布。经过正态性检验和偏度计算，R统计量呈正偏态分布，l_{zp}则是呈负偏态分布。表 1 展示了这两种统计量在不同测验长度下，给定显著性水平性上的经验临界值。可以发现，l_{zp}在不同项目长度下的临界值相近，这是l_{zp}公式标准化建构的结果，而R的临界值会随着项目长度增加而不断增大，这是因为R是由多项非负加权残差累加得到，项目数量越多，R也会越大。

4.2 研究二：R统计量和l_{zp}的检测率

在获得了各种测验长度下，正常被试群体的 PFS 分布后，依靠设定不同的一类错误率，可以得到对应的临界值。研究二中的项目参数与研究一相同，区别在于加入了部分受异常因素影响的被试。

异常的测验行为可能以多种方式表现出来，例如作弊，题目预知，低测验动机等

(Meijer & Sijtsma, 2001; Rupp, 2013)。参考 Karabatso(2003)、Doval 和 Delicado(2020)的研究中关于异常被试的模拟条件，我们设定了几种多级计分下的异常作答情况，包括①作弊②幸运猜测③随机作答④粗心⑤创造性作答⑥混合。其定义和操作定义如表 2 所示。主要可以总结为以下几类：(1)低能力者的异常高能力表现(①②)；(2)高能力者的异常低能力表现(④⑤)；(3)广泛存在的随机作答表现③。需要注意的是，实验中模拟的异常类型并不能够完全代表现实中可能出现的情况，主要目的是体现上述三种异常类型，因此对其命名和定义的解读应当谨慎。

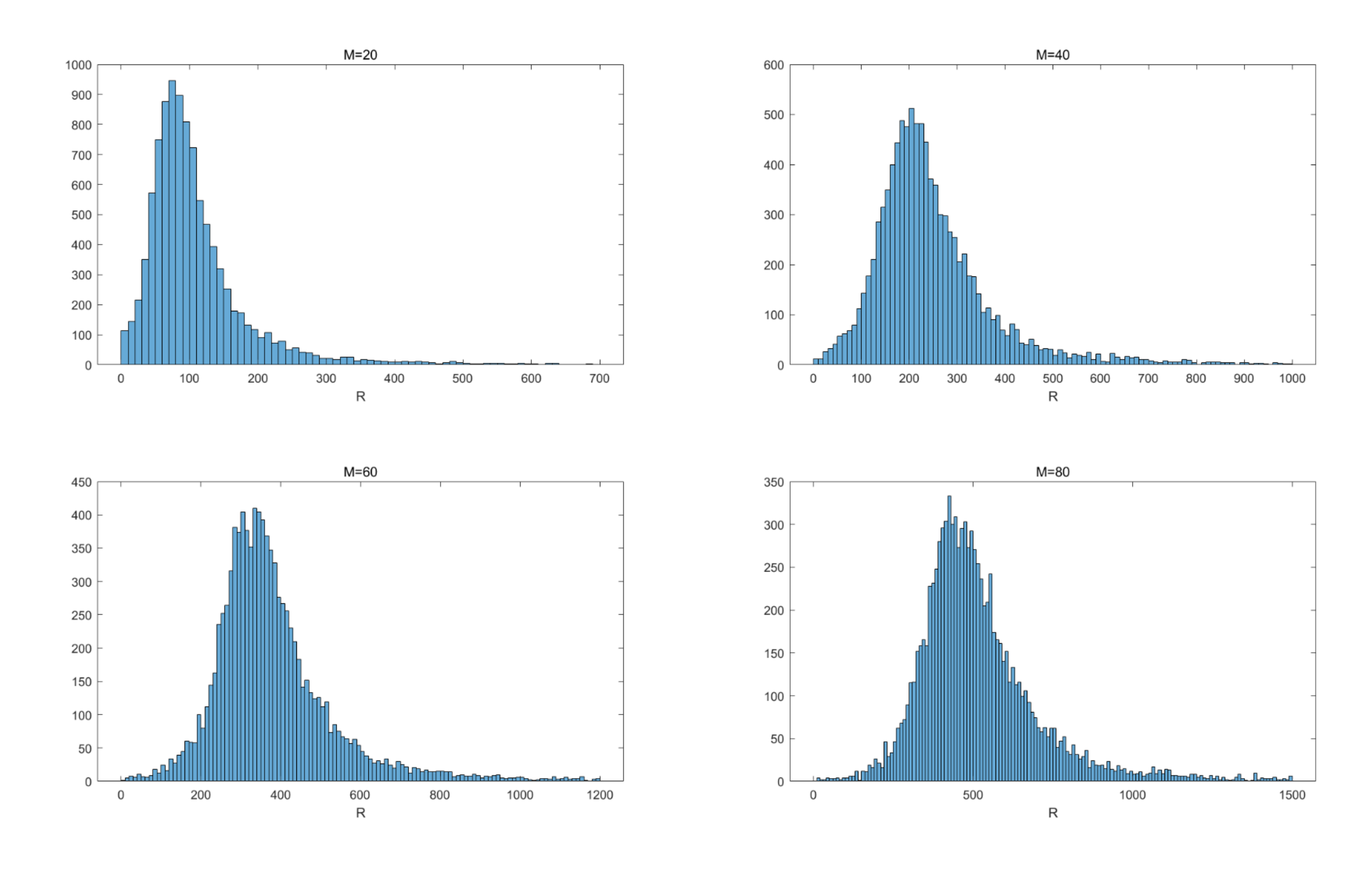


图 1 *R*统计量在项目长度为 20，40，60，80 个项目下的分布

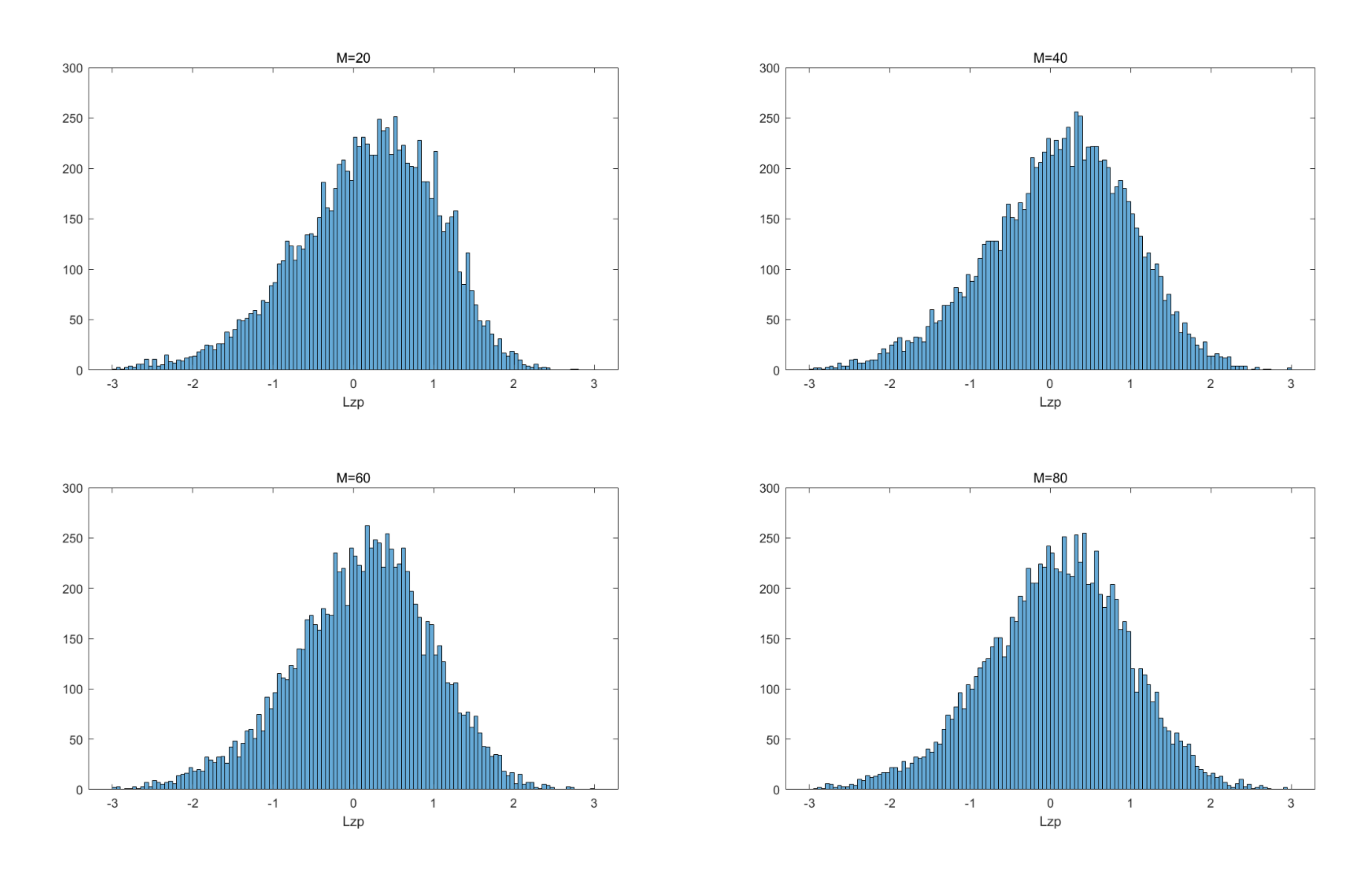


图 2 l_{zp}统计量在项目长度为 20，40，60，80 个项目下的分布

表 1 l_{zp}和R在测验长度为 20，40，60，80 项目下的截断点值

α=0.05，α=0.025，α=0.01(单侧)

项目数	一类错误率	R(UB)	l_{zp}(LB)
20	0.01	706.9	-2.215
	0.025	416.2	-1.770
	0.05	282.9	-1.399
40	0.01	1057.4	-2.176
	0.025	691.9	-1.760
	0.05	519.6	-1.417
60	0.01	1407.7	-2.125
	0.025	949.2	-1.717
	0.05	730.4	-1.383
80	0.01	1904.7	-2.127
	0.025	1278.3	-1.738
	0.05	983.6	-1.411

注：UB、LB 代表 upper bound 和 lower bound，即上限值和下限值。

被试的异常作答项目数$m = p \times M$，M为测验长度。将异常程度p作为自变量引入实验中，分别等于 0.1，0.25，0.5(代表低程度，中等程度，高程度)，整个研究共有 4×6×3=72 种实验条件组合，每种实验条件下模拟 3000 名被试参加测验，重复 100 次。对于异常行为在考生中的“流行程度”，我们参考已有研究(Curran et al., 2010; Meade &Craig, 2012; Rupp, 2013; Meade, 2016; Shao et al., 2016; Yu & Cheng, 2019)，将异常被试在群体中的占比设定为 20% 。

由于本研究的主要目的是比较两种统计量对于异常作答行为的检测能力，故考察在项目参数已知的情况下它们的表现，这样异常被试的多少不会对项目参数产生影响，考生因异常行为导致的能力—作答不拟合将会更好地通过 PFS 表现出来。一旦被试的 PFS 超过了临界值($l_{zp} < l'_{zp}$，$R > R'$)，就将该被试标记为异常。检测率的定义是标记为异常的被试占异常被试的比例，同时，虚警率也将在结果中标注出来，其值为错误标记的正常被试比例。由于项目参数不会随着异常被试人数的变化而产生影响，虚警率会相当接近临界值对应的一类错误率。实验结果如下表 3 到表 6 所示。

从表 3 到表 6 我们可以发现，实际的虚警率和事先设定的一类错误率十分接近，这是因为临界值是从正常被试群体中选取的，所以该群体分布中仍然有部分极端个案会超过临界值，导致被错误判定为异常，理论上该部分被试的比例与临界值对应的一类错误率是一致的。

表 2 异常的测验行为定义及其操作定义

异常类型	定义	操作定义
作弊	能力较低的被试在平均难度较高的项目上获得满分	随机挑选低能力被试($\theta < Z.375$),在难度最高的前n个项目上获得满分
幸运猜测	能力较低的被试在平均难度较高的项目上依靠猜测获得满分	随机挑选低能力被试($\theta < Z.375$)，在难度最高的前n个项目上，有 0.2 的概率获得满分，0.8 的概率维持原作答
随机作答	所有能力范围内的被试都有可能出现，有一定概率获得 0 分	随机挑选被试，随机抽取n题，有 0.8 的概率得 0 分，0.2 的概率维持原作答
粗心	能力较高的被试在平均难度较低的项目上有一定概率获得 0 分	随机挑选高能力被试($\theta > Z.625$)，在难度最低的前n个项目上，有 0.8 的概率获得 0 分，0.2 的概率维持原作答
创造性作答	能力较高的被试在最容易的项目上获得 0 分	随机挑选高能力被试($\theta > Z.625$)，在难度最低的前n个项目上获得 0 分
混合	将以上异常情况进行混合	以上五种情况各占异常被试总体的五分之一

注：$\theta < Z.375$表示能力由低到高排序的前37.5%，$\theta > Z.625$表示能力由低到高排序的后37.5%。

一方面，在测验长度为 20，异常程度低(即受异常作答行为影响的题目百分比为 10%)时，R对各类异常行为的检验力都比 l_{zp}要高，平均高出 17.6%；而当异常程度达到中等或高(即受异常作答行为影响的题目百分比为 25%和 50%)时，两者的平均检测率几乎相同；在测验长度为 40，异常程度低时，R对各类异常行为的检验力平均高出l_{zp}11.2%，在异常程度达到中和高时，二者的平均检测率相当接近；当测验长为 60 时，低异常程度时，R对各类异常行为的检验力平均高出l_{zp}8.5%；当测验长度达到 80 时，低异常程度时，R对各类异常行为的检验力平均高出l_{zp}5.7%。可以看出，随着测验长度的增加，低异常程度时，R对各类异常行为的检测率与l_{zp}的差距在变小。在中和高异常程度下，两种统计量的平均检测率相当接近，一些情况下，R略占优。

另一方面，两种统计量对于不同类型异常作答模式表现出了不同的检测率，比如都对检测创造性作答和作弊的表现较好，通常都在 90%以上，不同的是在作弊且低异常程度下，l_{zp}性能显著弱于R。随着异常程度的上升，l_{zp}会稍微领先，这是因为虽然R具有对异常高能力的敏感性，但也缺乏对异常程度变化的稳健性。另外，它们对于探测其它类型异常行为的检测力相对不高，这是因为作弊和创造性作答属于较为极端的异常类型，理论较容易被检测出，而其他类型迷惑性较高，相对更难发现。并且两种统计量体现出了一些共同的趋势，比如检测率会随着异常题目覆盖率的提高而提高，也会随着测验长度的增加而提高。

表 3 测验长度为 20 个项目时R和l_{zp}的一类错误率和检测率

异常类型	临界值对应一类错误率	异常程度低(0.1)				异常程度中(0.25)				异常程度高(0.5)			
		虚警率		检测率		虚警率		检测率		虚警率		检测率	
		R	l_{zp}	R	l_{zp}	R	l_{zp}	R	l_{zp}	R	l_{zp}	R	l_{zp}
作弊	0.01	0.010	0.011	0.499	0.224	0.009	0.011	0.592	0.937	0.010	0.010	0.714	0.994
	0.025	0.024	0.026	0.729	0.400	0.023	0.026	0.827	0.973	0.024	0.026	0.901	0.998
	0.05	0.049	0.052	0.889	0.581	0.047	0.051	0.957	0.989	0.048	0.052	0.972	1
幸运猜测	0.01	0.010	0.011	0.124	0.031	0.009	0.011	0.230	0.096	0.010	0.011	0.457	0.295
	0.025	0.025	0.026	0.196	0.067	0.024	0.026	0.362	0.165	0.024	0.026	0.579	0.396
	0.05	0.049	0.052	0.262	0.119	0.048	0.052	0.472	0.243	0.048	0.052	0.673	0.487
随机作答	0.01	0.010	0.010	0.138	0.068	0.010	0.010	0.181	0.205	0.010	0.011	0.173	0.387
	0.025	0.025	0.025	0.201	0.120	0.025	0.025	0.272	0.279	0.025	0.026	0.321	0.465
	0.05	0.050	0.050	0.270	0.185	0.050	0.050	0.363	0.354	0.051	0.051	0.463	0.535
粗心	0.01	0.010	0.011	0.826	0.491	0.009	0.011	0.832	0.887	0.009	0.011	0.646	0.995
	0.025	0.024	0.026	0.914	0.632	0.025	0.026	0.952	0.934	0.024	0.026	0.907	0.998
	0.05	0.050	0.051	0.946	0.736	0.051	0.052	0.985	0.961	0.050	0.052	0.989	0.999
创造性作答	0.01	0.009	0.011	0.922	0.704	0.010	0.011	0.839	0.998	0.009	0.011	0.653	1
	0.025	0.024	0.026	0.989	0.854	0.025	0.026	0.966	1	0.025	0.026	0.957	1
	0.05	0.050	0.052	0.999	0.939	0.051	0.052	0.995	1	0.050	0.051	0.997	1
混合	0.01	0.010	0.010	0.459	0.299	0.010	0.011	0.464	0.600	0.010	0.012	0.430	0.659
	0.025	0.025	0.025	0.552	0.405	0.025	0.026	0.585	0.638	0.025	0.028	0.596	0.688
	0.05	0.050	0.051	0.615	0.495	0.051	0.051	0.661	0.669	0.051	0.054	0.678	0.715

表 4 测验长度为 40 个项目时R和l_{zp}的一类错误率和检测率

异常类型	临界值对应一类错误率	异常程度低(0.1)				异常程度中(0.25)				异常程度高(0.5)			
		虚警率		检测率		虚警率		检测率		虚警率		检测率	
		R	l_{zp}	R	l_{zp}	R	l_{zp}	R	l_{zp}	R	l_{zp}	R	l_{zp}
作弊	0.01	0.009	0.011	0.589	0.326	0.009	0.011	0.905	1	0.009	0.011	0.799	1
	0.025	0.022	0.027	0.794	0.516	0.023	0.027	0.991	1	0.022	0.027	0.954	1
	0.05	0.046	0.053	0.925	0.676	0.046	0.053	1	1	0.046	0.053	0.997	1
幸运猜测	0.01	0.009	0.011	0.140	0.025	0.009	0.011	0.522	0.193	0.009	0.011	0.591	0.432
	0.025	0.023	0.026	0.205	0.059	0.023	0.027	0.617	0.292	0.023	0.026	0.733	0.552
	0.05	0.047	0.054	0.274	0.107	0.046	0.053	0.682	0.389	0.046	0.052	0.822	0.653
随机作答	0.01	0.010	0.010	0.127	0.072	0.010	0.010	0.197	0.295	0.010	0.010	0.168	0.438
	0.025	0.025	0.025	0.187	0.130	0.025	0.026	0.291	0.378	0.025	0.025	0.329	0.510
	0.05	0.050	0.050	0.255	0.199	0.049	0.050	0.393	0.454	0.050	0.051	0.487	0.575
粗心	0.01	0.010	0.011	0.576	0.472	0.010	0.010	0.891	0.982	0.011	0.010	0.677	0.995
	0.025	0.025	0.026	0.770	0.624	0.026	0.026	0.979	0.992	0.025	0.025	0.920	0.998
	0.05	0.050	0.051	0.898	0.735	0.050	0.051	0.995	0.996	0.050	0.050	0.985	0.999
创造性作答	0.01	0.010	0.010	0.639	0.695	0.010	0.010	0.886	1	0.010	0.010	0.813	1
	0.025	0.025	0.026	0.840	0.828	0.025	0.026	0.987	1	0.025	0.025	0.973	1
	0.05	0.050	0.052	0.952	0.906	0.050	0.051	0.999	1	0.050	0.050	0.994	1
混合	0.01	0.010	0.010	0.372	0.317	0.010	0.012	0.549	0.637	0.010	0.012	0.488	0.668
	0.025	0.025	0.025	0.503	0.423	0.025	0.029	0.620	0.663	0.025	0.029	0.617	0.701
	0.05	0.049	0.051	0.595	0.510	0.051	0.055	0.660	0.689	0.051	0.056	0.688	0.731

表 5 测验长度为 60 个项目时R和l_{zp}的一类错误率和检测率

异常类型	临界值对应一类错误率	异常程度低(0.1)				异常程度中(0.25)				异常程度高(0.5)			
		虚警率		检测率		虚警率		检测率		虚警率		检测率	
		R	l_{zp}	R	l_{zp}	R	l_{zp}	R	l_{zp}	R	l_{zp}	R	l_{zp}
作弊	0.01	0.009	0.010	0.930	0.790	0.009	0.011	0.975	1	0.009	0.011	0.868	1
	0.025	0.024	0.025	0.996	0.904	0.023	0.026	1	1	0.024	0.026	0.998	1
	0.05	0.047	0.050	1	0.957	0.048	0.052	1	1	0.048	0.051	1	1
幸运猜测	0.01	0.009	0.011	0.298	0.053	0.009	0.010	0.513	0.230	0.009	0.010	0.590	0.484
	0.025	0.023	0.026	0.366	0.105	0.023	0.026	0.618	0.341	0.023	0.026	0.722	0.605
	0.05	0.047	0.051	0.426	0.176	0.047	0.051	0.701	0.445	0.048	0.051	0.820	0.702
随机作答	0.01	0.010	0.009	0.275	0.162	0.010	0.010	0.375	0.429	0.010	0.010	0.430	0.655
	0.025	0.025	0.024	0.343	0.248	0.025	0.025	0.461	0.514	0.025	0.025	0.576	0.721
	0.05	0.050	0.049	0.414	0.333	0.050	0.050	0.549	0.590	0.050	0.049	0.696	0.775
粗心	0.01	0.010	0.011	0.968	0.881	0.009	0.011	0.987	0.998	0.009	0.010	0.974	1
	0.025	0.024	0.026	0.975	0.929	0.023	0.027	0.995	0.999	0.024	0.027	0.994	1
	0.05	0.048	0.052	0.981	0.956	0.047	0.052	0.999	1	0.048	0.052	0.999	1
创造性作答	0.01	0.010	0.010	1	0.999	0.009	0.011	1	1	0.010	0.011	1	1
	0.025	0.024	0.026	1	1	0.023	0.026	1	1	0.024	0.027	1	1
	0.05	0.048	0.053	1	1	0.047	0.052	1	1	0.048	0.052	1	1
混合	0.01	0.010	0.011	0.602	0.563	0.010	0.011	0.660	0.683	0.010	0.013	0.658	0.756
	0.025	0.025	0.026	0.625	0.612	0.025	0.028	0.679	0.710	0.026	0.030	0.724	0.776
	0.05	0.051	0.052	0.645	0.649	0.053	0.055	0.703	0.736	0.053	0.057	0.755	0.791

表 6 测验长度为 80 个项目时R和l_{zp}的一类错误率和检测率

异常类型	临界值对应一类错误率	异常程度低(0.1)				异常程度中(0.25)				异常程度高(0.5)			
		虚警率		检测率		虚警率		检测率		虚警率		检测率	
		R	l_{zp}	R	l_{zp}	R	l_{zp}	R	l_{zp}	R	l_{zp}	R	l_{zp}
作弊	0.01	0.010	0.011	1	0.963	0.009	0.011	0.985	1	0.009	0.011	0.985	1
	0.025	0.024	0.027	1	0.986	0.024	0.027	1	1	0.024	0.027	1	1
	0.05	0.048	0.052	1	0.995	0.047	0.053	1	1	0.047	0.053	1	1
幸运猜测	0.01	0.009	0.011	0.316	0.082	0.010	0.011	0.504	0.284	0.010	0.011	0.620	0.677
	0.025	0.024	0.026	0.400	0.148	0.024	0.026	0.639	0.404	0.024	0.027	0.780	0.776
	0.05	0.048	0.052	0.490	0.227	0.048	0.052	0.749	0.515	0.048	0.053	0.889	0.844
随机作答	0.01	0.010	0.010	0.178	0.133	0.010	0.010	0.242	0.406	0.010	0.010	0.195	0.561
	0.025	0.025	0.025	0.250	0.207	0.025	0.025	0.344	0.481	0.025	0.024	0.387	0.625
	0.05	0.050	0.050	0.325	0.288	0.050	0.050	0.450	0.548	0.049	0.049	0.562	0.683
粗心	0.01	0.009	0.011	0.962	0.912	0.009	0.011	0.954	1	0.009	0.011	0.797	1
	0.025	0.023	0.027	0.991	0.953	0.023	0.027	0.996	1	0.024	0.027	0.980	1
	0.05	0.046	0.053	0.997	0.975	0.046	0.053	1	1	0.047	0.052	0.999	1
创造性作答	0.01	0.009	0.011	0.999	0.995	0.009	0.011	0.960	1	0.009	0.011	0.956	1
	0.025	0.023	0.026	1	0.999	0.023	0.027	0.999	1	0.023	0.027	0.999	1
	0.05	0.047	0.053	1	1	0.046	0.053	1	1	0.046	0.053	1	1
混合	0.01	0.010	0.011	0.600	0.595	0.010	0.012	0.589	0.677	0.010	0.014	0.568	0.724
	0.025	0.025	0.026	0.616	0.623	0.025	0.029	0.633	0.707	0.026	0.032	0.644	0.752
	0.05	0.051	0.053	0.635	0.651	0.052	0.055	0.673	0.733	0.053	0.060	0.704	0.775

与此同时，我们还将l_{zp}与R在检测异常行为被试的 ROC 曲线下面积(area under curve，AUC)作为 PFS 的一种综合评价指标。因为 AUC 不依赖于固定阈值，表示了统计量的总体检测能力，其值越接近 1，说明该检测方法的性能就越好。通过比较R与l_{zp}在多种情境下的 AUC 值，可以用来评价统计量的性能。实验结果如下。

表 7 R和l_{zp}的 AUC 值

异常类型	异常程度	*M*=20		*M*=40		*M*=60		*M*=80	
		R	l_{zp}	R	l_{zp}	R	l_{zp}	R	l_{zp}
作弊	低(0.1)	0.979	0.921	0.986	0.937	0.997	0.989	0.999	0.996
	中(0.25)	0.987	0.995	0.996	0.998	0.997	0.998	0.998	0.998
	高(0.5)	0.989	0.998	0.994	0.998	0.995	0.998	0.997	0.998
幸运猜测	低(0.1)	0.643	0.598	0.689	0.610	0.771	0.669	0.828	0.709
	中(0.25)	0.775	0.704	0.891	0.795	0.919	0.829	0.940	0.860
	高(0.5)	0.888	0.827	0.961	0.908	0.964	0.925	0.979	0.962
随机作答	低(0.1)	0.696	0.649	0.699	0.671	0.780	0.752	0.752	0.724
	中(0.25)	0.756	0.736	0.780	0.800	0.848	0.861	0.825	0.842
	高(0.5)	0.807	0.834	0.808	0.846	0.902	0.927	0.867	0.898
粗心	低(0.1)	0.973	0.930	0.979	0.940	0.996	0.988	0.998	0.993
	中(0.25)	0.993	0.989	0.996	0.997	0.999	0.998	0.998	0.998
	高(0.5)	0.990	0.998	0.990	0.998	0.998	0.998	0.994	0.998
创造性作答	低(0.1)	0.997	0.985	0.987	0.980	0.999	0.998	0.999	0.998
	中(0.25)	0.994	0.998	0.996	0.998	0.999	0.998	0.997	0.998
	高(0.5)	0.991	0.998	0.993	0.998	0.998	0.998	0.997	0.998
混合	低(0.1)	0.827	0.794	0.824	0.803	0.845	0.837	0.846	0.838
	中(0.25)	0.854	0.849	0.854	0.859	0.872	0.877	0.873	0.877
	高(0.5)	0.866	0.870	0.865	0.876	0.885	0.891	0.880	0.888

为了更好地比较R与l_{zp}之间检测效能的差异，我们将二者的输出结果做差，图 3 中纵坐标表示二者检测率之间的差值，图 4 中纵坐标表示二者 AUC 之间的差值。由于R和l_{zp}的检测率在不同测验长度条件下趋势较为一致，因此我们仅选取了测验长度为 20 个项目的条件下二者之间的差值作图。

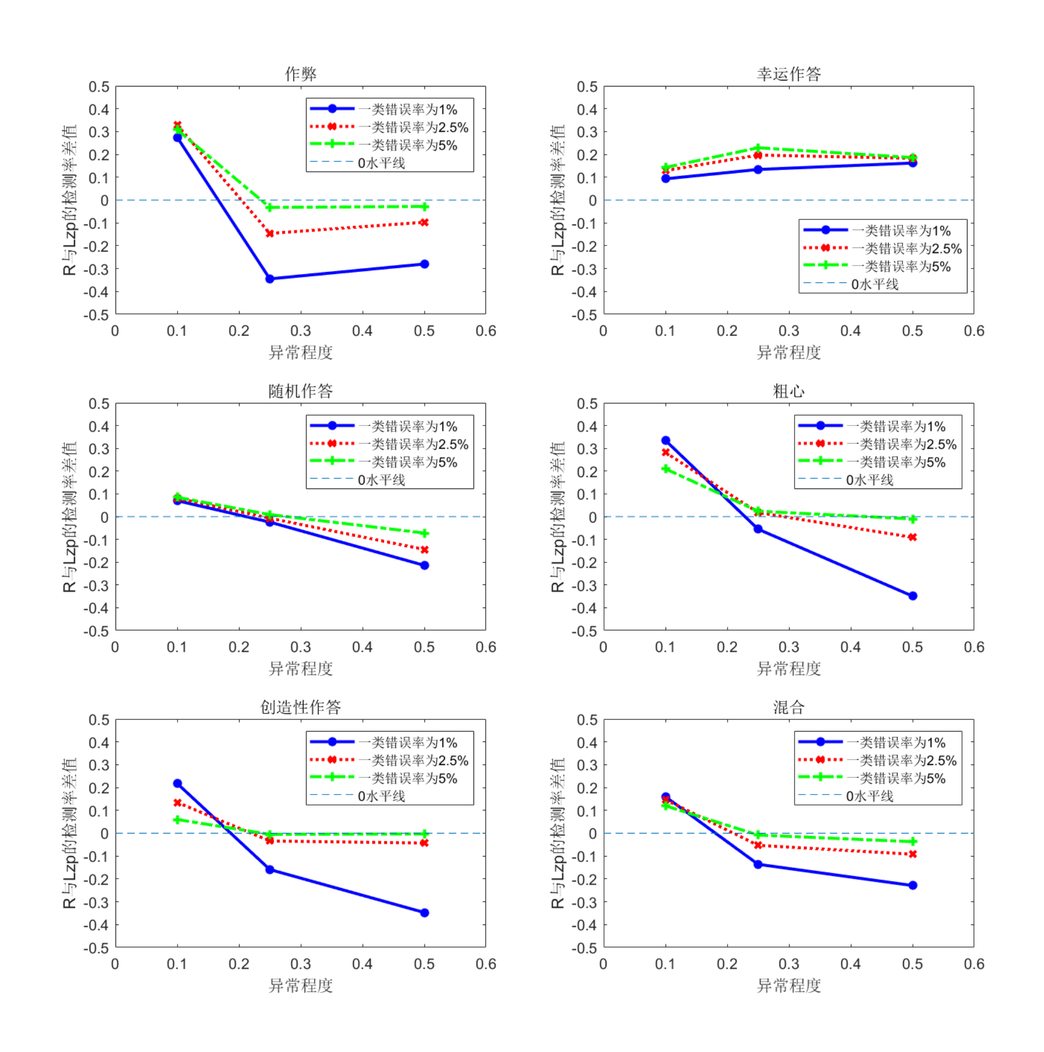


图 3 R和l_{zp}在测验长度为 20 个项目时检测率的差值

从图 3 中可以发现，在异常程度较低时，几乎所有条件下R的检测率均高于l_{zp}，而在异常类型为幸运猜测时，R的检测率全面高于 l_{zp}。这与前文中对R的分析结果基本一致。但随着异常程度的增加，l_{zp}逐渐占据优势，此时R的检测率接近或低于l_{zp}，这也体现了相较于l_{zp}，R更容易受到掩蔽效应变化带来的影响。需要注意的是，虽然绝大部分异常程度较低情况下，R相较于l_{zp}拥有更大的优势，但在测验长度较大时(测验包含 60, 80 个项目)，混合异常类型情况下R的检测率会略低于l_{zp}。在测验长度较小时(测验包含 20, 40 个项目)，R统计量在幸运猜测、随机作答和粗心的异常类型下，检测率会高于l_{zp}。而在其他情况下，R统计量的检测率会接近或略低于l_{zp}。

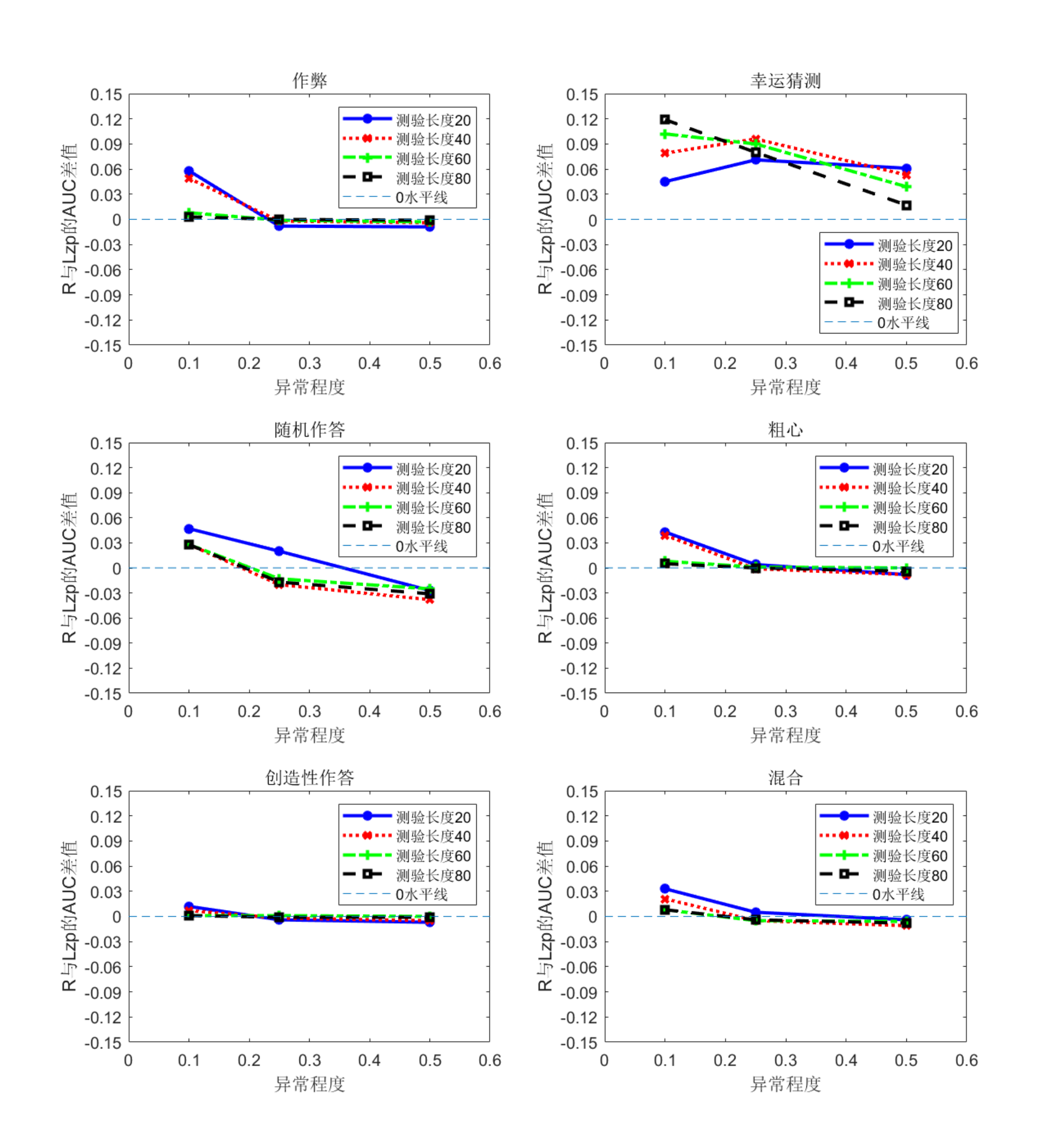


图 4 R和l_{zp} AUC 的差值

对于另一评价指标 AUC，从图 4 可以发现，除了异常行为类型为随机作答，异常程度为中高时，R的 AUC 值低于l_{zp}，其余情况下，R统计量的 AUC 值几乎全面高于或等于l_{zp}。由于 AUC 代表了 PFS 的综合性能，不受一类错误率设定的限制，这说明R的综合检测效果总体上好于l_{zp}，特别是在实验条件为幸运猜测这种典型的异常高能力表现的情况下，R的 AUC 明显高于l_{zp}。

5 实证研究

为了探究R在实际研究中的应用，我们使用的数据来自 1994 年到 1995 年的美国青少年健康纵向研究(national longitudinal study of adolescent health, NLSAH)，这是一项针对 7-12 年级青少年的教育和健康相关研究(Harris & Udry, 2010)，本研究所使用的数据集可以从

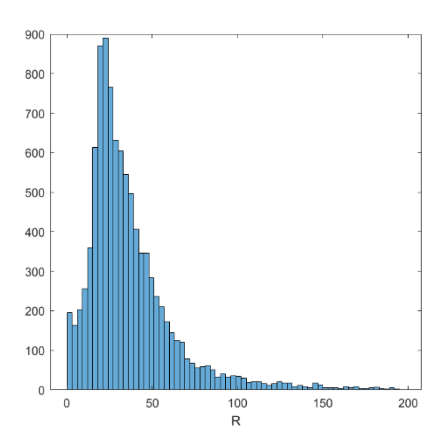
https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/21600/datadocumentation 获得。Yu 和 Cheng(2019)对该数据进行了探索性因素分析，结果支持该问卷仅考查一个维度的假设。该数据中包含 19 个用于测量被试情绪状态的项目，以四等级李克特量表的形式呈现。被试需要在四个选项中选取最符合自己过去一周内状态的描述，分别编码为 0，1，2，3。由于几乎没有被试选取类别四，即大多数时候符合或总是符合，所以类别 3 和类别 4 合并为了一个类别。

在筛选出存在缺失数据的 97 名被试后，将余下的 6457 名被试的数据用作分析。步骤一，我们使用 PARSCALE 4.1 软件工具，以 GRM 模型对作答数据进行拟合，目的是为得到 19 个项目的项目参数。鉴于已将类别 3、类别 4 合并为一个类别，所以最终存在 3 个选项类别，也即每个项目中存在 2 个项目难度参数(b_1, b_2)和 1 个区分度参数(a)，拟合结果如下。

表 8 全国青少年健康纵向研究(1994-1995)拟合 GRM 模型项目参数

题号	a	b_1	b_2	题号	a	b_1	b_2
1	0.859	0.393	2.199	11	0.759	-0.544	1.262
2	0.655	0.668	2.474	12	0.661	0.336	2.142
3	0.889	0.741	2.547	13	0.962	0.445	2.251
4	0.499	-0.778	1.028	14	0.729	0.740	2.546
5	0.814	-0.345	1.461	15	0.652	-0.155	1.651
6	1.136	0.272	2.078	16	1.314	0.064	1.870
7	0.813	-0.211	1.595	17	0.863	0.596	2.402
8	0.503	-1.028	0.778	18	0.827	0.023	1.829
9	0.917	1.424	3.230	19	0.859	1.794	3.600
10	0.831	0.946	2.752	-	-	-	-

步骤二中，前文已经提到，要使用R和l_{zp}，需要生成一批正常考生的得分数据，来获得特定项目条件下的分布，以此来确定临界值。在软件 Matlab 2020a 中模拟生成 10000 名能力服从标准正态分布的被试(6457 名考生的能力接近正态分布，均值为-0.0003，标准差 0.9222)，让这批被试按照 GRM 的作答概率进行作答模拟，计算这 10000 名被试的R和l_{zp}。为了使实验条件接近真实情况，计算零分布时所使用的被试能力参数均为估计值。根据前文的研究结果，选取 5%的一类错误率截断点，可以做到控制一类错误率较小的情况下，获得尽可能高的检测率。R和l_{zp}的分布如下图所示。R在该分布下的第 95 百分位数为 98.49，l_{zp}在该分布下的第 5 百分位数为-1.23；



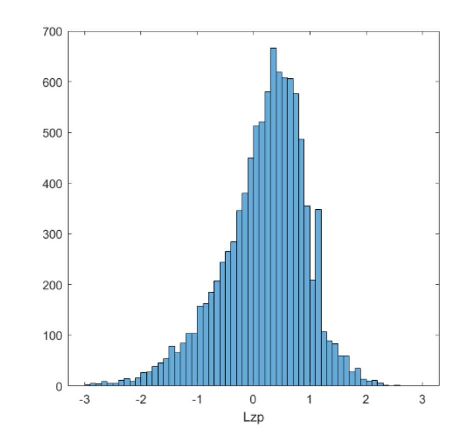


图 5 美国青少年健康纵向研究(1994-1995)模拟 10000 名被试的R和l_{zp}分布

步骤三，计算原始数据中 6457 名被试的R和l_{zp}，以和临界值进行比较，来判断被试是否存在异常行为。经过比较得到，6457 名被试中，由R标记出的异常行为被试有 423 人，占比 6.6%，由l_{zp}标记出的异常行为被试有 562 人，占比 8.7%，同时违反R和l_{zp}判定标准的被试有 253 人。由于实证研究中，被试是否真实存在异常行为是未知的，因此无法计算虚警率以及检测率。考虑到模拟研究的结果显示，l_{zp}通常相较于R更为严苛，即更容易将被试判断为异常，所以本研究中由l_{zp}标记出的异常行为被试数量相对较多也符合预期。

为了更一步考察两个统计量的表现，我们对被检测出的考生的作答进行分析。分别选取了三类根据统计量标记为异常的被试共 15 人的作答模式，如下表所示

表 9 标记为异常的被试作答模式

类型	被试序号	作答向量
共同	15	(2,1,0,0,0,0,0,0,0,0,0,2,0,0,2,0,0,0,0)
	23	(0,1,0,2,1,0,0,1,2,0,2,2,1,0,2,2,2,0,2)
	48	(2,0,2,2,1,2,0,2,2,1,2,1,2,2,2,2,2,0,2)
	49	(1,0,0,2,0,0,0,0,0,0,0,0,0,0,0,1,0,0,2)
	50	(2,2,2,0,2,2,2,1,0,0,1,2,0,0,1,0,0,0,0)
R	43	(0,0,2,2,0,0,1,1,0,1,1,0,0,0,0,0,0,1,0)
	45	(0,0,0,2,1,2,0,0,0,1,0,0,0,0,0,1,0,0,0)
	99	(1,0,0,1,0,1,1,0,0,2,1,2,0,0,0,0,1,1,0)
	108	(0,0,0,1,0,0,0,2,0,0,1,0,0,0,2,2,0,0,0)
	114	(0,2,0,0,1,0,0,0,0,0,0,0,0,0,2,0,0,0,0)
l_{zp}	6	(1,0,0,1,2,0,2,0,1,0,2,1,0,0,1,0,1,0,1)
	36	(0,0,2,2,1,2,0,2,1,0,2,1,2,0,2,1,0,0,0)
	52	(1,0,2,2,2,0,2,0,1,2,2,1,2,0,2,2,2,1,0)
	55	(2,2,2,1,2,2,1,2,0,2,1,0,2,0,1,2,0,1,0)
	101	(0,1,0,1,0,2,1,2,1,2,2,0,1,0,2,1,1,1,1)

注：“共同”表示同时被R和l_{zp}标记为异常，“R”表示仅被R标记为异常，“l_{zp}”表示仅被l_{zp}标记为异常

可以发现：(1)同时被两种统计量标记异常的被试，都具有较为明显的异常状况，得分较为极端，如第 15、49 号被试，其估计能力较低，然而他们都在不止一个项目上作答为 2，不符合低能力者的正常作答模式。再如第 48 号被试，其在大量项目上作答为 2，估计能力应该偏高，但是却在第 2、7、18 题上作答为 0，也同样不符合高能力者应有的作答模式；(2)仅被R标记的异常被试群体，具有较为明显的特点，即作答中包含大量的 0，这说明这部分被试以低能力者居多，但是却能在少数项目上获得 2，属于异常高能力表现。R“专注”于将这部分被试筛选出来，也与前文中介绍的R的特点相一致；(3)对于仅被l_{zp}标记的被试，他们的作答结果较为“均匀”，这可以理解为不同于R对低能力者的异常高能力表现的特异性，l_{zp}对不同异常类型检测的覆盖面更广。

6 讨论和未来的研究方向

本文提出了一种基于残差的个人拟合统计量R，其在检测低能力被试获得异常高能力表现方面具有一定的优势。在多级计分模型的背景下，比较了R和l_{zp}的经验分布(通常R呈现正偏态分布，l_{zp}则呈现负偏态分布)，随着测验长度的增加，二者分布的偏度逐渐降低。

在模拟研究中，设置不同的一类错误率，获得R和l_{zp}的经验临界值，用以区分正常被试和存在异常的被试。通过模拟被试在测验中不同程度以及不同类型的异常作答行为，使用R和l_{zp}，计算比较被试的个人拟合指标和临界值的相对位置，来判断被试是否异常，用检测率和 AUC 作为R和l_{zp}检测异常被试的指标。需要注意的是，尽管研究二中列举了三种大小的一类错误率(0.01，0.025，0.05)及其对应的检测率，但在实际应用中，我们需要进行权衡，力求控制一类错误率的情况下，尽可能提高检测率，研究结果显示，选取 0.05 的一类错误率最为合适。因此，我们着重讨论了 0.05 一类错误率水平下二者检测率的差异。研究结果显示：中低异常程度时(即异常测验行为覆盖的项目数较低)，R相对 l_{zp}拥有更好的检测效果，特别是在异常行为类型为幸运猜测(低能力被试有较大概率获得高分)时，R的检测效果相较于l_{zp}具有显著优势，这与理论分析中R具备异常高能力敏感性的特点相一致；而在其他条件下，二者则较为接近，或l_{zp}略优于R。考虑到l_{zp}在同条件下的实际一类错误率通常会略高于R，这将导致其检测率略微地升高。同理，个人拟合指标的一类错误率在模拟研究中通常存在一定的膨胀，理论检测率应当略小于表格所展示的值。对于不涉及一类错误率的 AUC 指标，R的优势会更加地明显。

在实际测验情景中，R具有较高的应用价值，这是因为在大部分选拔测试中，低能力者更倾向于主动产生异常作答行为，来获得与高测验成绩相挂钩的切身利益，其行为通常

会对测验造成损害，如作弊行为伴随的项目泄露等；相比之下，高能力者的异常作答行为偏向被动，危害性也相对较低。Rupp(2013)的文献中，系统地回顾了前人关于异常测验行为的研究，其中低能力者表现出的作弊和猜测行为相对更受到研究者的关注，这也说明作弊和猜测行为是众多测验异常行为中影响较为广泛，研究者迫切需要解决的问题。实际使用当中，由于个人拟合统计量的有效性依赖于项目参数，当项目参数未知时，在包含“异常作答数据”的基础上进行的参数估计会产生“遮蔽效应”，因此这种情况下最好对数据进行适当的处理，比如考虑稳健的参数估计方法(Cooperman et al., 2021) 或者进行数据清理(Hong et al., 2020)等。并且，考虑到R的特性，可以将R与其他统计量结合使用，以发挥最好的效果。

本文在五级计分的背景下进行了模拟实验，在实证研究部分基于已知项目参数在三级计分下进行了模拟实验，其结果和五级计分下基本一致。在未来可以考虑在更多其他等级计分的情况下进行研究验证。另外，由于R使用的是经验临界值，在一定程度上会影响它的使用。未来的研究方向可以是在R的基础上，对其进行拓展，得到其标准化的形式，使其分布为正态分布或渐进正态分布，临界值的选取就不必依赖于特定的测验项目参数，省去获取临界值的步骤；或是在残差处理中进行改进，使其保留初步分辨异常测验行为类型的信息；又或者对R进行拓展，使其适用于更为广泛的 IRT 模型。最后，本研究中未涉及项目参数对于两种统计量的影响，为了考虑更为广泛的实际应用场景，将项目参数的估计误差纳入考虑也是未来需要进一步深入考虑的问题。

参考文献

Buchanan, T., & Smith, J. L. (1999). Using the internet for psychological research: Personality testing on the world wide web. *British Journal of Psychology, 90(*1), 125–144.

Chen, Q., Ding, S., Zhu, L., & Xu, Z. (2010). Three-parameter graded response model and its parameter estimation. *Journal of Jiangxi Normal University (Natural Science)*, *34*(2), 117–122.

[陈青, 丁树良, 朱隆尹, 许志勇. (2010). 三参数等级反应模型及其参数估计. *江西师范大学学报(自然科学版), 34*(2), 117–122.]

Cheng, X., Ding, S., Zhu, L., & Wu, H. (2012). The stratified item selection strategy with maximal information under graded response model. *Journal of Jiangxi Normal University (Natural Science)*, *36*(5), 117–122.

[程小扬, 丁树良, 朱隆尹, 巫华芳. (2012). 等级评分模型下的最大信息量分层选题策略. *江西师范大学学报(自然科学版), 36*(5), 446–451.]

Cooperman, A. W., Weiss, D. J., & Wang, C. (2021). Robustness of adaptive measurement of change to item parameter estimation error. *Educational and Psychological Measurement*, Advance online publication.

Curran, P. G., Kotrba, L., Denison, D. (2010, April). *Careless responding in surveys: Applying traditional techniques to organizational settings.* Paper presented at the 25th annual conference of the Society for Industrial/Organizational Psychology, Atlanta, GA.

Dodd, B, G., Ayala, R, J, De., & Koch, W, R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*(1), 5–22.

Doval, E., & Delicado, P. (2020). Identifying and classifying aberrant response patterns through functional data analysis. *Journal of Educational and Behavioral Statistics, 45*(6), 719–749.

Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement*, *28*(1), 105–113.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67–86.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*(1), 59–79.

de la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45*(2), 159–177.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement, 32*(3), 224–247.

Fung, W. K. (1993). Unmasking outliers and leverage points: A confirmation. *Journal of the American Statistical Association, 88*(422), 515–519.

Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika, 72*(2), 159–180.

Gulliksen, H. (1950). *Theory of mental tests*. John Wiley & Sons Inc.

Guttman, L. (1944). A basis for scaling qualitative data. *American sociological review*, *9*(2), 139–150.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, et al., *Measurement and Prediction* (60–90). Princeton: Princeton University Press.

Harris, K. M., & Udry, J. R. (2010). *National Longitudinal Study of Adolescent Health (Add Health), 1994 –2008: Core files* [restricted use] (Technical report). Ann Arbor, MI: Inter-University Consortium for Political and Social Research.

Hong, Steedle, & Cheng. (2020). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement, 80* (2), 312–345.

Hotaka, Maeda. (2017). *Robust latent ability estimation based on item response information and model fit* (Dissertation). Milwaukee.

Huang, J. L., Bowling, N. A., Liu, M. Q., & Li, Y. H. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology, 30*, 299–311.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277–298.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*(4), 269–290.

Li, J., Ding, S. (2018). The several stratified methods of CAT in the presence of calibration error on GRM. *Journal of Jiangxi Normal University (Natural Science)*, *42*(4), 374–378.

[李佳, 丁树良. (2018). 基于 GRM 模型的 CAT 分层方法在校准误差中的应用研究. *江西师范大学学报(自然科学版), 42*(4), 374–378.]

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice, 26*(4), 29–37.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden (Ed.),

Handbook of modern item response theory (pp. 101–121). New York, NY: Springer.

Meade, A. W. (2016). Understanding and detecting careless responding in survey research. Retrieved from https://cba.unl.edu/outreach/carma/documents/CARMA-Meade-Presentation.pdf.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455.

Meijer, R., & Sijtsma, K. (2001). Methodology review, evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107–135.

Meijer, R., & Sijtsma, K. (2001). Dection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education, 8*(3), 261–272

Nering, M, L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19*(2), 121–129.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*(3), 200–219.

Rogers, H. J., & Hattie, J. A. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement, 11*, 47–57

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in Item Response Theory, Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling, 55*(1), 3–38.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34*(4), 1–97.

Schnipke, D. L. (1996). How contaminated by guessing are item-parameter estimates and what can be done about it? *Paper presented at the annual meeting of the National Council on Measurement in Education*, New York, NY.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213–232.

Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis.

Psychometrika, *81*(4), 1118–1141.

Sinharay, S. (2016). Asymptotically correct standardization of person-fit statistics beyond dichotomous items. *Psychometrika, 81*(4), 992–1013.

Snijders, T. (2001). Asymptotic distribution of person-fit statistics with estimated person parameter. *Psychometrika, 66*(3), 331–342.

van Der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement, 25*(3), 273–282.

van Krimpen-Stoop, M, L, A., & Meijer, R, R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement,* 26(2), 164–180.

Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement, 40*(4), 307–330.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design. Rasch measurement.* Chicago: Mesa Press.

Xiong, J., Luo, H., Wang, X., & Ding, S. (2018). The online calibration based on graded response model. *Journal of Jiangxi Normal University (Natural Science), 42*(1), 62–66.

[熊建华, 罗慧, 王晓庆, 丁树良. (2018). 基于 GRM 的在线校准研究. *江西师范大学学报(自然科学版), 42*(1), 62–66.]

Xiong J., Ding S., Luo F., & Luo Z. (2020) Online calibration of polytomous items under the graded response model. *Frontiers in Psychology. 10*:3085, 1–11.

Yu, X., & Cheng, Y. (2019). A change-point analysis procedure based on weighted residuals to detect back random responding. *Psychological Methods*, *24*(5), 658–674.

Yuan, K. H., & Zhong, X. (2008). Outliers, leverage observations, and influential cases in factor analysis: Using robust procedures to minimize their effect. *Sociological Methodology, 38*(1), 329–368.

Detection of aberrant response patterns using a residual-based statistic in testing with polytomous items

TONG Hao[1], YU Xiaofeng[1], Qin Chunying[2], PENG Yafeng[1], ZHONG Xiaoyuan[1]
([1]School of Psychology, Jiangxi Normal University, Nanchang, 330022, China)([2]School of Mathematics and Information Science, Nanchang Normal University, Nanchang 330032, China)

Abstract

Tests are widely used in educational measurement and psychometrics, and the examinee's aberrant responses will affect the estimation of their abilities. These examinees with aberrant responses should not be treated with conventional methods, the important thing is to accurately screen them out of the normal group. To achieve this, a common method is to construct person-fit statistics to detect whether the response patterns fit their estimated abilities.

In this study, a residual-based person-fit statistic R was proposed, which can be applied to both dichotomous or polytomous IRT models. The construction of R is based on a weighted residual between the observed response and the expected response. By accumulating the weighted residuals, the goodness of fit can be calculated and compared with a specific critical value to determine whether an examinee is aberrant or not. Given that tests with polytomous items can provide more information, polytomously scored items are being increasingly popular in educational measurement and psychometrics. The ability of R statistic to detect aberrant response patterns under the graded response model was mainly considered in this article.

An existing polytomous person-ft statistic l_{zp} was also introduced in its outstanding standardized form and superior power. In the first study, a simulation study was conducted to generate the empirical distribution of R statistic and l_{zp}. R statistic is an accumulation of weighted residuals, showing a positive skew distribution; l_{zp} shows a negative skew distribution when the test is less than 80 items. Both of them differ from the standard normal distribution, It is necessary to set critical value according to the type 1 error, using it to distinguish whether each respondent's response pattern is fitted. In the second study, examinees with different aberrant behaviors (e.g., Cheaters, Lucky guessers, Random respondents, Careless respondents, Creative respondents and Mixed) under different test length conditions were simulated, and the detection rate as well as area under curve (AUC) were used to compare the effectiveness of the two person-fit statistics. The results show that the R statistic has a better detection rate than l_{zp} when the aberrant behavior affects only a few items or the aberrant behavior is cheating or guessing. When the aberrant behavior covers plenty of items, l_{zp} is slightly better than R statistic. Then, an empirical study was also conducted to show the power of R statistic.

The optimize range of R statistic and l_{zp} are different, so we may combine R statistic and l_{zp} in the future person-fit studies. R statistic has a better detection rate under certain conditions compared to l_{zp}, especially when cheating and lucky guessing happened. Considering that cheating and guessing behaviors of low-ability examinees are more preferred in many aberrant test behaviors, the R statistic is valuable for further research and exploration in real-world applications.

Keywords: appropriateness measurement, item response theory, residual-based person-fit statistic, aberrant detection, polytomous item response models

附录 A：五级计分下，*R*和υ的比较

表 A-1 测验长度为 40 个项目时*R*和υ的一类错误率和检测率

异常类型	临界值对应一类错误率	异常程度低(0.1)				异常程度中(0.25)				异常程度高(0.5)			
		虚警率		检测率		虚警率		检测率		虚警率		检测率	
		R	υ	*R*	υ	*R*	υ	*R*	υ	*R*	υ	*R*	υ
作弊	0.05	0.053	0.061	1	1	0.050	0.059	1	1	0.045	0.053	1	1
幸运猜测	0.05	0.044	0.065	0.523	0.270	0.048	0.061	0.862	0.658	0.054	0.059	0.970	0.907
随机作答	0.05	0.054	0.052	0.718	0.570	0.054	0.050	0.662	0.618	0.049	0.055	0.747	0.720
粗心	0.05	0.057	0.059	0.623	0.335	0.047	0.055	0.855	0.697	0.049	0.059	0.958	0.898
创造性作答	0.05	0.051	0.059	1	1	0.048	0.056	1	1	0.051	0.053	1	1
混合	0.05	0.053	0.054	0.850	0.740	0.043	0.041	0.973	0.887	0.043	0.058	0.975	0.963

表 A-2 测验长度为 80 个项目时*R*和υ的一类错误率和检测率

异常类型	临界值对应一类错误率	异常程度低(0.1)				异常程度中(0.25)				异常程度高(0.5)			
		虚警率		检测率		虚警率		检测率		虚警率		检测率	
		R	υ	*R*	υ	*R*	υ	*R*	υ	*R*	υ	*R*	υ
作弊	0.05	0.053	0.048	1	1	0.056	0.049	1	1	0.052	0.050	1	1
幸运猜测	0.05	0.044	0.045	0.743	0.395	0.053	0.045	0.955	0.815	0.049	0.046	0.992	0.988
随机作答	0.05	0.045	0.048	0.557	0.570	0.053	0.040	0.807	0.722	0.047	0.047	0.770	0.737
粗心	0.05	0.046	0.050	0.797	0.388	0.043	0.053	0.987	0.797	0.055	0.045	0.995	0.988
创造性作答	0.05	0.047	0.045	1	1	0.055	0.045	1	1	0.049	0.059	1	1
混合	0.05	0.053	0.046	0.935	0.787	0.051	0.048	0.960	0.943	0.049	0.049	0.987	0.998